A NEW CONCEPT OF SKILL SCORE FOR RATING QUANTITATIVE FORECASTS

EDWARD M. VERNON

Weather Bureau Forecast Center, San Bruno Calif.

[Manuscript received November 8, 1951]

ABSTRACT

Skill scores for rating quantitative forecasts are proposed to take into account the deviations occurring between forecast and observed values. One score, the "deviation" skill score, weights the forecasts linearly according to the deviation; a second score, the "quadratic" skill score weights them according to the square of the deviation. These two scores are compared with the conventional skill score for two sets of forecasts, and for the same forecasts with bias introduced. It is concluded that use of either the deviation or the quadratic skill score is preferable to use of the conventional skill score in rating quantitative forecasts. Examples of the step-by-step computations of the two new scores are given.

THE DEVIATION SKILL SCORE

The skill score, as first proposed by Heidke [1] and used during recent years for certain forecast verification purposes, may be written

$$S=\frac{R-E}{T-E} \tag{1}$$

where S is the skill score, R the number of correct forecasts, T the total number of forecasts, and E the number of forecasts expected to be correct on some standard such as chance.

This method of computing a skill score places the same weight on each incorrect forecast regardless of the amount by which the observed condition deviates from the forecast. In other words, a deviation of say 10 class intervals has no more effect on the skill score than one of but 1 class interval. For some purposes it would be advantageous to have the skill score evaluate the actual amount by which forecast and observed conditions differ, i. e., take into account the magnitude of error. To accomplish this end, an analogous equation for skill score may be written

$$S_d=\frac{\sum d_e-\sum d_f}{\sum d_e} \tag{2}$$

where S_d is the skill score which considers magnitude of deviations, hereafter referred to as the "deviation skill score," Σd_f is the sum of deviations occurring between forecast and observed values, and Σd_e is the sum of deviations to be expected on some basis such as chance.

The value of Σd_f and Σd_e can best be expressed in terms of row, column, and cell totals in the typical contingency table, wherein the frequencies of forecast values are arrayed in columns and of observed values in rows, while a given cell is identified by the row and column to which it alone is common. When the standard of reference is chance, the summations become

$$\sum d_f=\sum(n_{rc}d_{rc}) \tag{3}$$

$$\sum d_e=\sum\left(\frac{n_r n_c}{T}\, d_{rc}\right) \tag{4}$$

where n_r is the number of cases falling in a given row; n_c is the number of cases falling in a given column; n_{rc} is the number of cases in the cell at the intersection of row r and column c; $n_r n_c/T$ is the number which would have fallen by chance in the cell representing the intersection of row r and column c; d_{rc} is the deviation represented by that cell and is equal to the number of class intervals by which the cell is removed from the perfect hit cell for the same column.

When the standard of reference is climatological expectancy, according to one of the more common definitions of that standard, Σd_f remains as expressed in (3) while Σd_e becomes

$$\sum d_e=\sum\left(\frac{n_r n_{cc}}{T}\, d_{rc}\right) \tag{5}$$

where n_{cc} represents the climatological expectancy for the column, i. e., the number of times which climatological averages would lead one to expect the observed conditions to fall in the particular class interval represented by the column. The other symbols in (5) remain as previously defined in (4) and (1).

THE QUADRATIC SKILL SCORE

In the foregoing equations all deviations are weighted linearly, a deviation of one class interval scoring as one unit deviation, two class intervals as two unit deviations, and so on. Where it is desired to have the penalty increase as the square of the deviation this is accomplished by substituting d_f^2 for d_f; d_e^2 for d_e; and d_{re}^2 for d_{re}. Hence, the quadratic deviation skill score* S_{d2}, in the computation of which the penalty increases as the square of the deviation of forecast from observed conditions, becomes

$$S_{d2}=\frac{\Sigma d_e^2-\Sigma d_f^2}{\Sigma d_e^2} \tag{6}$$

Where the standard of comparison is chance, Σd_f^2 and Σd_e^2 are given by

$$\Sigma d_f^2=\Sigma(n_{re}d_{re}^2) \tag{7}$$

and,

$$\Sigma d_e^2=\Sigma\left(\frac{n_r n_c}{T}\,d_{rc}^2\right) \tag{8}$$

But if the standard of comparison is climatological expectancy, as previously defined in connection with equation (5), Σd_e^2 becomes:

$$\Sigma d_e^2=\Sigma\left(\frac{n_r n_{cc}}{T}\,d_{re}^2\right) \tag{9}$$

and Σd_f^2 remains as expressed in (7).

COMPARISON OF SKILL SCORES

Both the deviation skill score as computed by (2), and the quadratic skill score as computed by (6), conform to the usual conception of a skill score in that they vary on a scale of from zero to 1, with a value of zero indicating complete lack of skill over the standard of comparison, usually chance or climatological expectancy; and a value of 1 indicating the highest possible skill, with all observed data falling in the forecast class intervals. It is apparent that all three of these skill scores, S, S_d, and S_{d2} will be identical when the forecasts are either perfect or completely without skill. Just how they compare for the vast majority of forecasts which fall between these two extremes can be visualized to some extent by comparing scores attained on two sets of forecasts (A and B) presented in table 1.

TABLE 1.—*Contingency tables of precipitation forecasts by A and B, and corresponding S, S_d, and S_{d2}*

Forecasts by A

Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65	5	2	1	1	74
0.01–0.20	9	6	3	3	1	22
0.21–0.50	4	5	4	2	1	16
0.51–1.00	1	2	3	3	1	10
≥1.01	1	1	2	2	2	8
Total	80	19	14	11	6	80 / 130

S=0.362
S_d= .498
S_{d2}= .602

Forecasts by B

Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65	9				74
0.01–0.20	10	6	6			22
0.21–0.50	5	4	4	2	1	16
0.51–1.00			4	3	3	10
≥1.01				6	2	8
Total	80	19	14	11	6	80 / 130

S=0.362
S_d= .639
S_{d2}= .824

In these two examples, based on hypothetical data, precipitation forecasts are made for the amount of rain which will fall. Rain is forecast and recorded in five class intervals as indicated in the column and row headings. The data for forecasts by A and for those by B have been arranged so that each forecaster scores the same number of direct hits, namely 80. This together with the fact that their row and column totals are identical causes both to attain the same skill score (0.36) as computed in the conventional way by equation (1). However, it is clear that if we attach any significance to the amount by which the forecast is missed, forecasts by B were superior to those by A. Forecaster B had 19 fewer large misses than did Forecaster A, i. e., misses of 2, 3, and 4 class intervals. He had a proportionately larger number of near misses, i. e., misses of but one class interval. Now, if we use equation (2) and compute their deviation skill scores we find that B attains a higher value, scoring 0.64 against 0.50 for A. Furthermore, we note that if we square the deviations and compute quadratic skill scores by equation (6), the difference between the two sets of forecasts becomes even more pronounced, B scoring 0.82 against only 0.60 for A.

It would appear that for verifying quantitative forecasts, the deviation and quadratic skill scores, as herein defined, give better indications of the relative degree of skill than does the conventional skill score. However, before reaching such a conclusion we must consider the possibility that rating on the basis of the size of the deviations will lead forecasters to bias their forecasts by forecasting the middle class interval, where the largest possible deviation is at a minimum, rather than trying to catch extreme conditions by forecasting the extreme class intervals where the largest possible deviation is at a maximum.

To examine this possibility, the forecasts by A and B have been biased by placing all of the rain forecasts in the middle class interval as shown in table 2. In other words, every time A forecasts rain we have placed the forecast in

TABLE 2.—*Contingency tables of biased precipitation forecasts by A and B, and corresponding scores S, S_d, and S_{d2}*

Forecasts by A

Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65		9			74
0.01–0.20	9		13			22
0.21–0.50	4		12			16
0.51–1.00	1		9			10
≥1.01	1		7			8
Total	80		50			77 / 130

S=0.323
S_d= .469
S_{d2}= .582

Forecasts by B

Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65		9			74
0.01–0.20	10		12			22
0.21–0.50	5		11			16
0.51–1.00			10			10
≥ 1 1.01			8			8
Total	80		50			76 / 130

S=0.310
S_d= .482
S_{d2}= .631

*Hereafter referred to as the quadratic skill score.

TABLE 3.—*Computation of "deviation" skill score, S_d, for set of forecasts appearing in table 3A below*

TABLE 3A.—*Array of frequencies of forecast and observed values:*

n_r, n_c, *and* n_{rc}

		Forecast					
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	
Observed	0	65	5	2	1	1	74
	0.01–0.20	9	6	3	3	1	22
	0.21–0.50	4	5	4	2	1	16
	0.51–1.00	1	2	3	3	1	10
	≥1.01	1	1	2	2	2	8
	Total	80	19	14	11	6	80 130

TABLE 3B.—*Array of deviation represented by each cell:*

d_{rc}

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	0	1	2	3	4
	0.01–0.20	1	0	1	2	3
	0.21–0.50	2	1	0	1	2
	0.51–1.00	3	2	1	0	1
	≥1.01	4	3	2	1	0

TABLE 3C.—*Array of number of cases expected by chance in each cell:*

$$\frac{n_r n_c}{T}$$

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	45.54	10.82	7.97	6.26	3.42
	0.01–0.20	13.54	3.22	2.37	1.86	1.02
	0.21–0.50	9.85	2.34	1.72	1.35	0.73
	0.51–1.00	6.15	1.46	1.08	0.85	0.46
	≥1.01	4.92	1.17	0.86	0.68	0.37

TABLE 3D.—*Array of chance frequencies weighted by cell deviation:*

$$\frac{n_r n_c}{T} d_{rc} = \text{(table 3B) (table 3C)}$$

$$\sum \left(\frac{n_r n_c}{T} d_{rc}\right) = \sum d_e = 155.26$$

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	0	10.82	15.94	18.78	13.68
	0.01–0.20	13.54	0	2.37	3.72	3.06
	0.21–0.50	19.70	2.34	0	1.35	1.46
	0.51–1.00	18.45	2.92	1.08	0	0.46
	≥1.01	19.68	3.51	1.72	0.68	0

TABLE 3E.—*Array of forecast frequencies weighted by cell deviation:*

$$n_{rc} d_{rc} = \text{(table 3A) (table 3B)}$$

$$\Sigma (n_{rc} d_{rc}) = \Sigma d_f = 78$$

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	0	5	4	3	4
	0.01–0.20	9	0	3	6	3
	0.21–0.50	8	5	0	2	2
	0.51–1.00	3	4	3	0	1
	≥1.01	4	3	4	2	0

$$S_d = \frac{\Sigma d_e - \Sigma d_f}{\Sigma d_e} = \frac{155.26 - 78}{155.26} = .498$$

TABLE 4.—*Computation of "quadratic" skill score, S_{d2}, for set of forecasts appearing in table 4A below*

TABLE 4A.—*Array of frequencies forecast and observed values:*

n_r, n_c, *and* n_{rc}

		Forecast					
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	
Observed	0	65	5	2	1	1	74
	0.01–0.20	9	6	3	3	1	22
	0.21–0.50	4	5	4	2	1	16
	0.51–1.00	1	2	3	3	1	10
	≥1.01	1	1	2	2	2	8
	Total	80	19	14	11	6	80 130

TABLE 4B.—*Array of squared deviation represented by each cell:*

d^2_{rc}

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	0	1	4	9	16
	0.01–0.20	1	0	1	4	9
	0.21–0.50	4	1	0	1	4
	0.51–1.00	9	4	1	0	1
	≥1.01	16	9	4	1	0

TABLE 4C.—*Array of number of cases expected by chance in each cell:*

$$\frac{n_r n_c}{T}$$

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	45.54	10.82	7.97	6.26	3.42
	0.01–0.20	13.54	3.22	2.37	1.86	1.02
	0.21–0.50	9.85	2.34	1.72	1.35	0.73
	0.51–1.00	6.15	1.46	1.08	0.85	0.46
	≥1.01	4.92	1.17	0.86	0.68	0.37

TABLE 4D.—*Array of chance frequencies weighted by squared cell deviation:*

$$\frac{n_r n_c}{T} d^2_{rc} = \text{(table 4B) (table 4C)}$$

$$\sum \left(\frac{n_r n_c}{T} d^2_{rc}\right) = \sum d^2_e = 387.4$$

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	0	10.82	31.88	56.34	54.72
	0.01–0.20	13.54	0	2.37	7.44	9.18
	0.21–0.50	38.40	2.34	0	1.35	2.92
	0.51–1.00	55.35	5.84	1.08	0	0.46
	≥1.01	78.72	10.53	3.44	0.68	0

TABLE 4E.—*Array of forecast frequencies weighted by square of cell deviation:*

$$n_{rc} d^2_{rc} = \text{(table 4A) (table 4B)}$$

$$\Sigma (n_{rc} d^2_{rc}) = \Sigma d^2_f = 154$$

		Forecast				
		0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
Observed	0	0	5	8	9	16
	0.01–0.20	9	0	3	12	9
	0.21–0.50	16	5	0	2	4
	0.51–1.00	9	8	3	0	1
	≥1.01	16	9	8	2	0

$$S_{d2} = \frac{\Sigma d^2_e - \Sigma d^2_f}{\Sigma d^2_e} = \frac{387.4 - 154}{387.4} = .602$$

the column representing a forecast of 0.21 to 0.50 inch in which the largest possible deviation is two class intervals. Forecasts by B were biased in the same manner. The verification of forecasts by A and B, biased in this manner, is shown in table 2. By comparing these scores with those accompanying table 1 we can see what effect the bias produced on the scores.

It is readily apparent that even on the poorer forecasts, i. e., those by A, there was some loss in score caused by the introduction of bias. The conventional skill score S dropped from 0.36 to 0.32, a loss of 4 points. S_d and S_{d2} dropped by smaller amounts, 3 and 2 points respectively. Thus it would at first appear that the conventional skill score places the greatest penalty on bias and is in that respect to be preferred over the deviation skill score and the quadratic skill scores presented in this article. However, before reaching such a conclusion let us see how the bias affected scores on the better set of forecasts, i. e. forecasts by B.

Here we see that while the conventional skill score S dropped 5 points because of the bias, the deviation skill score dropped 16 points and the quadratic skill score 19 points because of the same bias. Thus we see that for quantitative forecasts attaining a relatively high degree of skill in forecasting the correct class interval, the quadratic and deviation skill scores penalize unwarranted bias more than does the conventional skill score. However, on forecasts attaining a relatively low degree of skill the situation is reversed. This can be interpreted as an argument for use of either the deviation or the quadratic skill score in preference to the conventional skill score in rating quantitative forecasts, because their use would on the one hand appear to encourage forecasting the exact class interval where the verification expectancy is sufficiently high, and on the other hand would place a minimum penalty on biasing the forecast toward the middle class interval when the forecaster knows that his verification expectancy is low.

EXAMPLES OF COMPUTATIONS

To assist the reader in visualizing the application of the formulae, two sets of computations are shown in tables 3 A–E, and tables 4 A–E. These tables show how both the "deviation" and "quadratic" skill scores were computed for forecasts by A appearing in table 1.

REFERENCE

1. P. Heidke, "Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst," *Geografiska Annaler*, vol. 8, No. 4, 1926, pp. 310–349.